Wrangling efforts

1. I gathered data about dog tweets from three sources:
    - Manually, a CSV file (The WeRateDogs Twitter archive from Udacity)
    - Programmatically, a TSV file (via the requests library)
    - Twitter API (with the help from the Tweepy library)

2. I assessed data visually both visually and programmatically for quality and tidiness. I tried to look for completeness, validity, accuracy and consistency. When I assessed the data programmatically I mainly used pandas methods such as:
    - .head()
    - .tail()
    - .sample()
    - .info()
    - .value_counts()
    - And other various methods on indexing and selecting data.

    My quality findings are as following:
    - timestamp: wrong data type (string --> timestamp)
    - tweet_id: wrong data type (int --> string)
    - rating_numerator, rating_denominator ratings: ratings are not accurate / have different ranges and should be normed
    - name: many "none" names as well as lower case, incorrect words like "a" and "an"
    - p1, p2, p3: inconsistent spelling of dog types
    - retweeted_status_id: retweeted tweets are an issue, since we are just interested in the original tweets
    - reply_to_status_id: replied tweets are an issue, since we are just interested in the original tweets
    - source: too complicated, can be displayed easier for the 4 possible values
    - drop columns that are not used for analysis (retweet and reply columns)

    My tidiness findings are as following:
    - rating_numerator and rating_denominator can be combined into one column (for a better comparison)
    - 4 dog stage columns (doggo, floofer, ...) are categorical data and should thus be put into one column
    - tables could be merged via the tweet_id

3. I cleaned the data with regard to the teached “define-code-test” schema. Before I started I made a copy of the original data, so that they got untouched.
    - I started to change the timestamp data type to datetime
    - I merged all three data frames into one, based on their common tweet_id
    - I removed the retweet-related and the reply-related rows

- After having deleted the retweet-related and the reply-related rows I drop the respective columns
- I replaced all ‘None’ and np.NaN values from the "doggo", "floofer", "pupper", "puppo" columns and then merged the remaining values into the new ‘stage’ column. Furthermore there were cases where two values were existent, e.g. “Doggo” and “Pupper” both had values in their columns. Thus, I concatenated them with commas.
- I converted the tweet_id to a string
- I converted all dog types to lower_case
- I simplified the “source” values via extracting the display string from the ahref link
- I generated an additional column for getting a quotient from the rating. (rating-numberatior / rating-denomiator) * 10
- I replaced no-name-lowercase words and “None” values with NaN in the name column

4. I stored the data in a csv master file.